IEOR 142 Project Report
Group Members: Richard Vo, Mario Rincon, Rachael Lam, KC Kim, Ian Lin

Link to the Code: https://drive.google.com/file/d/1vP5qXI_PhxLiIGEK-D9vwSqmH9mmwbUE/view?usp=sharing
Link to the Data:

1. https://data.world/bgp12/nbancaacomparisons/workspace/file?filename=players.csv
2. https://www.kaggle.com/datasets/justinas/nba-players-data
3. https://www.kaggle.com/datasets/nathanlauga/nba-games?select=games_details.csv

Topic

Every year the NBA, National Basketball Association, hosts the NBA Draft to formally pick 75 players into their professional teams. 70 of those players come from the NCAA, National Collegiate Athletic Association, and around 4,000 are undrafted but still eligible to join a team. The NBA teams have a group of people that specialize in looking at the potentials who join the team. They have to go through thousands of statistics manually, which can be very grueling. We propose a model that can automate this process, and that will predict the best possible future stars that will boost the teams rosters. Creating this will reduce the amount of effort that each team has to put in to find the best pick that is right for them. Our model uses the data from NBA players who used to be NCAA players to predict how well they did in the NBA.

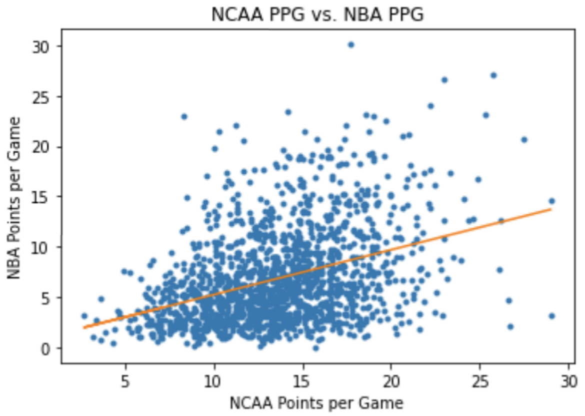
Data

We used a combination of seven datasets to help build our model. Our most used dataset is from data_world and contains content that is scraped from basketball reference and sports reference, websites that NBA teams use to source their players. The dataset contains features such as, physical measurements, NCAA point average, shooting percentages, and years in the association. The second dataset is taken from Kaggle and is similar to the first one but is more NBA specific. The last five datasets contain more specific information about the player's NBA statistics such as games, player's game details, recent nba players, rankings, and teams.

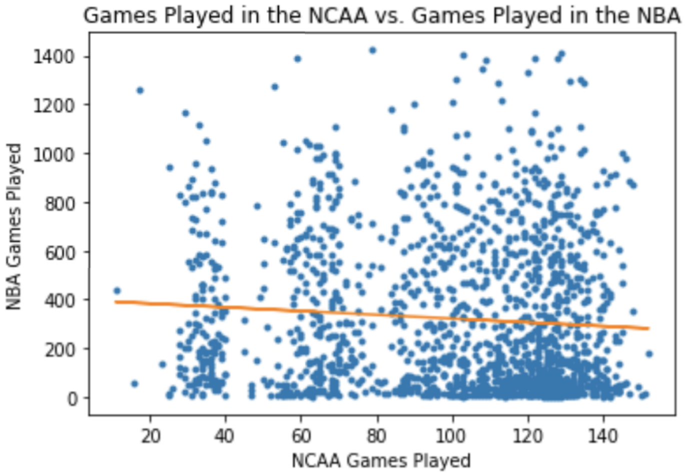
Cleaning

The dataset is generally pretty clean when it comes from an external source that specializes in datasets, but there were a couple things that we still needed to do. We started by dropping some columns that were not useful to our dataset, such as the first index unnamed column and url. While dropping all the missing values, we realized that effective field goal percentage had many of them, so we decided to also drop that column too. The last part of the cleaning also involved converting some of the numbers from strings to floats, so that it would be easier to use in EDA. We also renamed some of the columns so that they would be consistent in both the NBA and NCAA datasets. We also removed all the players who had a null value for their college name, because they did not participate in the NCAA. For some important missing values, we decided to use K-Nearest Neighbors to fill it in.

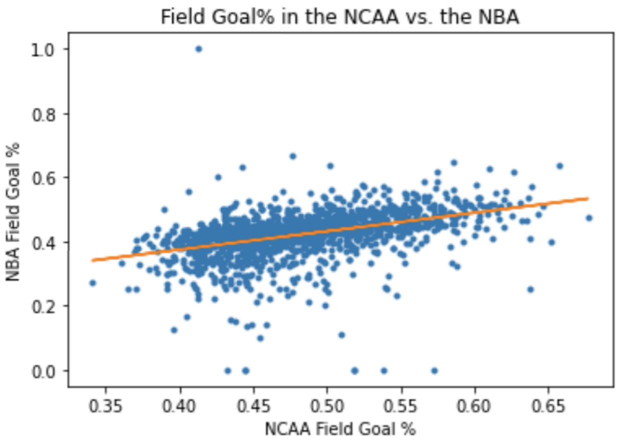
EDA and Visualizations



The goal of our EDA is to plot the similar variables that exist in both the NBA and the NCAA. We wanted to see which ones would be the most significant before plotting our whole model. We created a series of scatterplots, and the first one involved seeing the correlation between the PPG, points per game, for the player in both the NBA and the NCAA. There is a positive correlation between the two variables because the line is very linear. There are more clusters surrounding the x-axis (NCAA) which means that players were more likely to score in the NCAA. The points spread out more as they go along the y-axis. This is validated by the fact that the NBA pulls from the top players in the NCAA so therefore, it is much harder to score points. But, the positive correlation still shows that someone who scores many points in the NCAA will still score many points in the NBA.



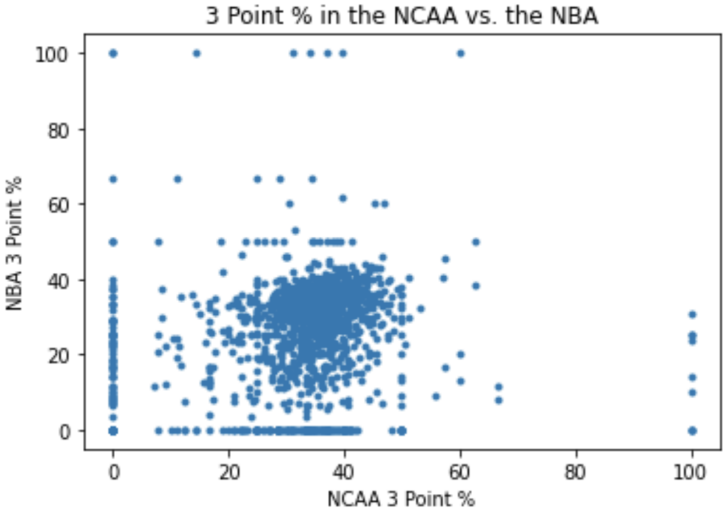
The next variable that we used a scatterplot on was how many games they played in both leagues. The plot seems to be plotted in three sections along the x-axis. The groups were 20-40, 50-80, and 90-140 games. The 20-40 games made sense because an NCAA season usually lasts 25-35 games, so some players only play one season. The top players in the league normally play one season in college to fulfill the NBA requirement before being drafted in the league. But the distribution of points along the y-axis seems to be the same no matter how far along the x-axis goes. This is consistent with my hypothesis that the best players get drafted after one season. Although many players get to play in the NBA after they finish their college careers, many of them do not get to play that many games. The largest cluster is where NCAA games are equal to 120-140 and NBA games are equal to 0-200.



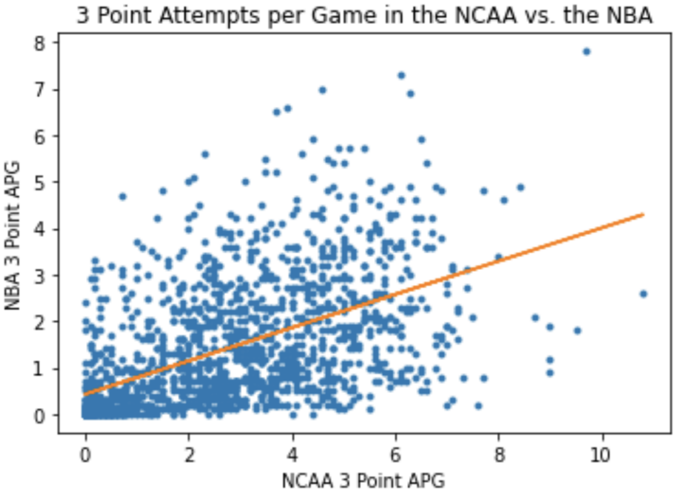
The next scatterplot shows the players field goal percentage. The plot is close to a horizontal line but still slightly linear, which means that all of the field goal percentages were around the same in the NBA, with an exception for a few outliers. Even people who had a lower field goal percentage in the

NCAA were able to have the same field goal percentage in the NBA as the high field goal percentage people. The biggest cluster is when the players have a 40% FG% in both the NBA and NCAA. This means that they are also looking for people who have room for improvement.

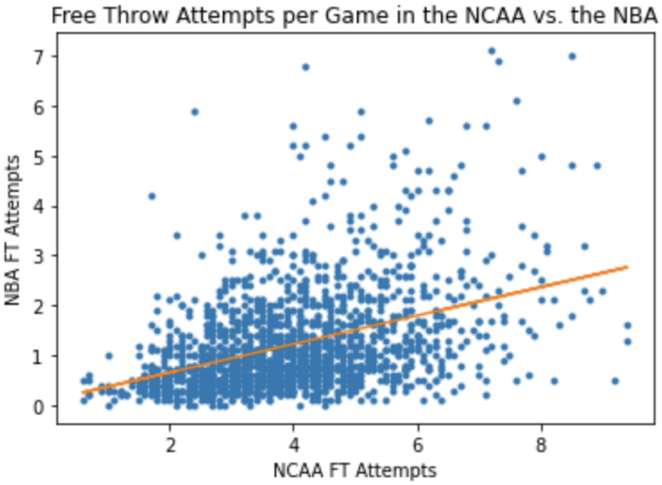
This plot shows the three point average percentage. The scatterplot is interesting because the points form a box, with a large cluster in the 30-40% range. Players who had previously scored 30-40% of their three pointers, normally had the same percentage in the NBA. But there was a very significant line along the x-axis, which meant that people who scored little to a lot of three pointers were not likely to attempt three pointers in the NBA.



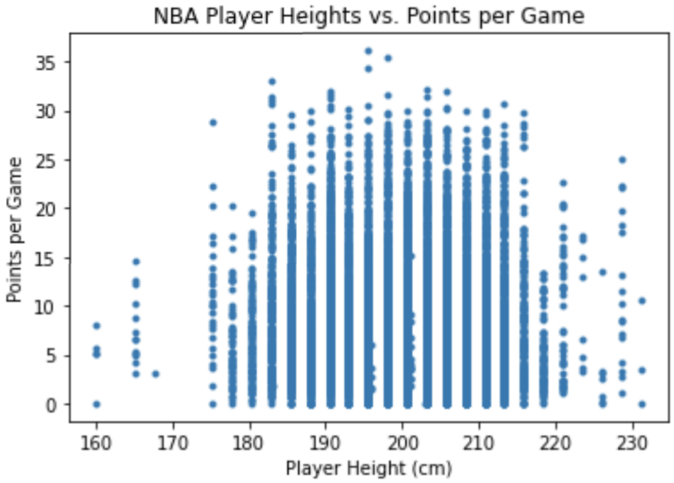
The next plot shows the three point attempts per game. The scatterplot is a lot more linear than the previous graph but it still shows similar trends. People were more likely to attempt three pointers in the NCAA than in the NBA. The biggest cluster is along the x - axis, where x is equal to 0-6 and y is equal to 0-2. Most players would attempt 0-6 three pointers in their NCAA games but the number decreases to 0-2 when they reach the NBA.



The next plot shows the free throw attempts per game. This plot also has a big cluster along the x axis. Many players attempted 0-6 free throws in the NCAA but attempted around 0-2 free throws in the NBA. After doing some research, we found an article done by the NCAA that explained this phenomenon. The article states that there is a trend where teams who won, scored a significant amount of free throws, so many teams try to get as many free throw attempts as possible.

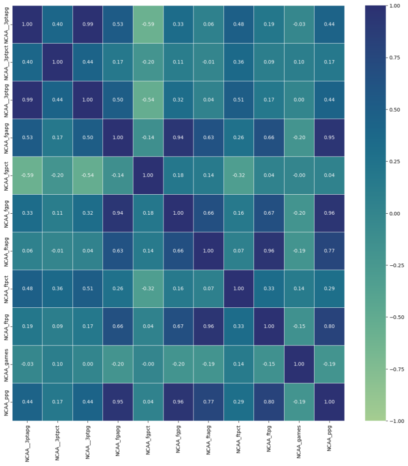


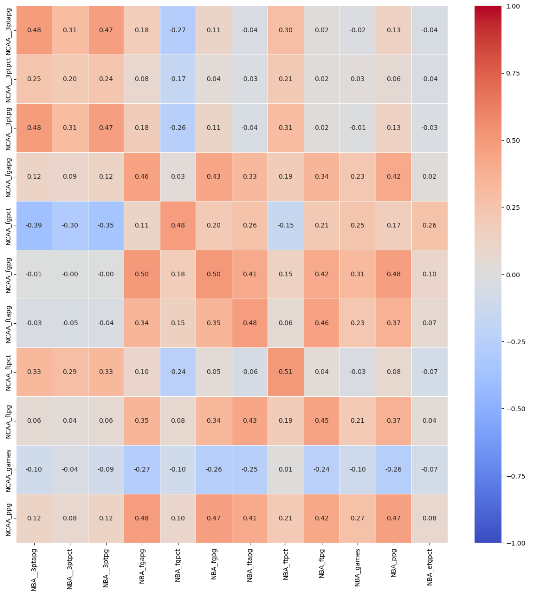
The next plot we had was nba players' heights vs how many points they scored in a game. Because the heights were so varied, it was hard to see a connection between the two variables. There was no correlation between the two, so height is not a good indicator of how many points a player will score in an NBA game.



Using a correlation matrix, we decided to plot some of the features in the NCAA dataset. Many of the features were highly correlated, such as, NCAA_fgapg and NCAA_fgpg, which indicates that multicollinearity exists in the dataset.

The next heat map plotted the same variables but it was NCAA vs. NBA. There is correlation between the same variables in the both datasets, as shown through the diagonal that runs from the top left to the bottom right. NBA_ftpct and NCAA_ftpct had the biggest correlation value out of all the values on the heatmap at a score of 0.51.

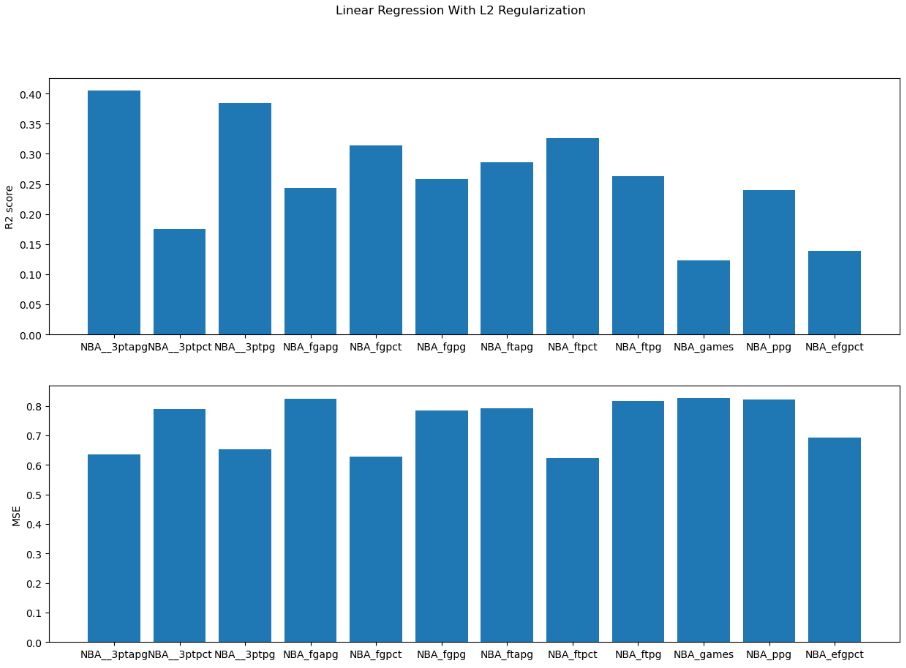




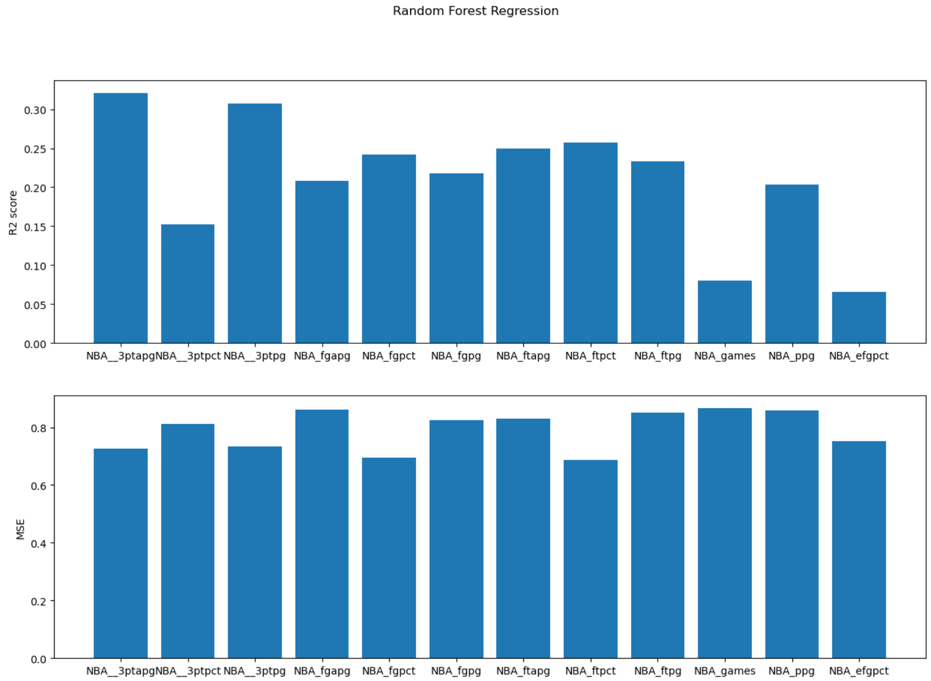
Model

We started by dividing the dataset into a training and test dataset. Because we are trying to predict how the player will do in the NBA, we are going to use the columns in the NBA dataset as our targets and the NCAA as our features. We tested a variety of models to see which one would perform the best by evaluating them by their r squared and mean squared error value.

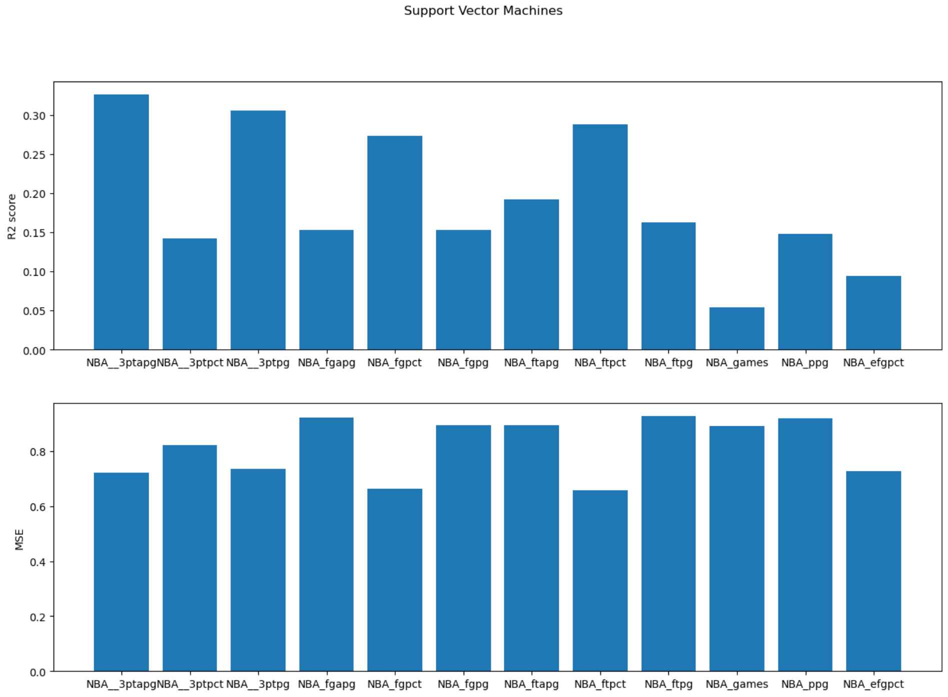
Because some of the plots in our EDA were linear, we decided to choose linear regression L2 regularization as our first model. The 3 points attempted per game had the highest r squared value of 0.41 out of all the variables, which aligns with the scatterplot we had, because it was quite linear. The field goals attempted per game and the numbers games the player played had the highest MSE value of 0.82. The overall r squared value for this model was 0.263 and the MSE was quite high at an average of 0.74



The next model we built was a Random Forest Regression model with cross validation to find the best value of the max_features. In this model also, three points attempted per game had the highest r squared value and the number of games that the player played had the highest mean squared value. The average r squared value was lower than the linear regression one at 0.212 and the average MSE was higher than the previous model at 0.791.

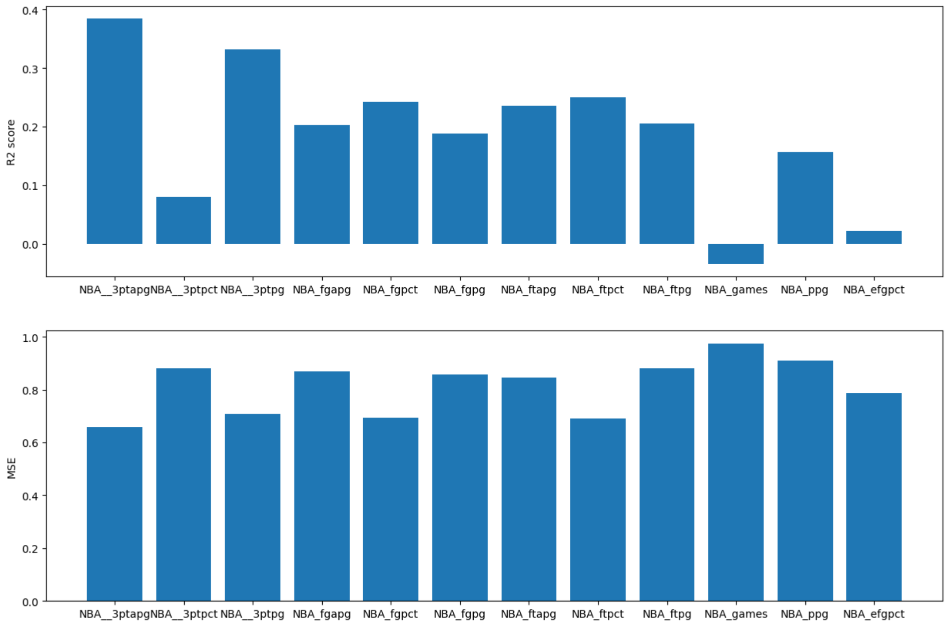

Random Forest Regression

The next model we built was a Support Vector Machine model with cross validation to find the best value of the regularization parameter c. In this model also, three points attempted per game had the highest r squared value but, the free throws per game had the highest MSE value. The average r squared value was lower at 0.191 and the average MSE was higher than the previous model at 0.815.

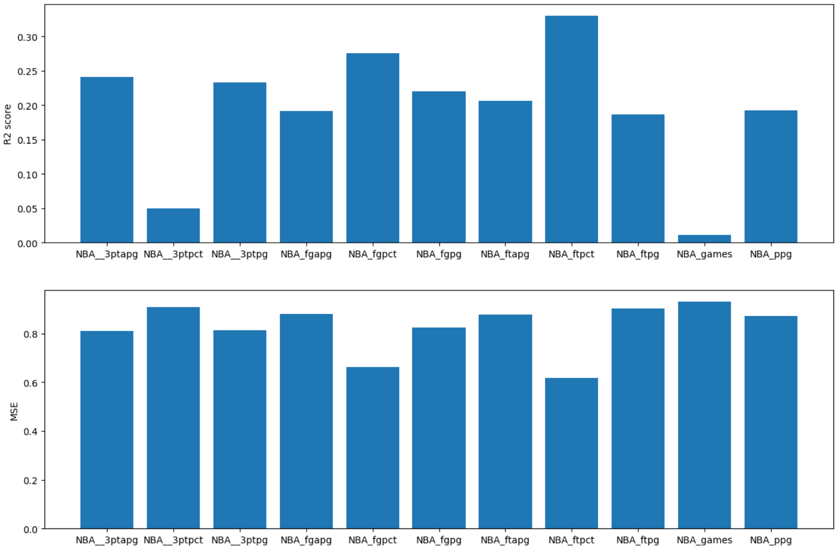

Support Vector Machines

The next model we built was a Neural Network with cross validation to find the best combination of hidden layer sizes. In this model, three points attempted per game had the highest r squared value and the number of games a player was in had the highest MSEA. The average r squared value was lower than the linear regression one at 0.189 and the average MSE was about the same as the previous model at 0.813.

Neural Network



The last model we built was a Linear Regression Model with 1D Input. The average r squared value was higher but similar to the other linear regression one at 0.194 and the average MSE was higher than the previous model at 0.828. Overall, the first Linear Regression Model did the best at predicting because of its r squared value. Because there was multicollinearity in the NCAA dataset, they were probably not the best at predicting and therefore most of the models did not have a high r squared value but performed very well when it came to MSE.



Linear Regression with 1D Input

Impact

Our work could have the potential to be useful to people who work in the sports industry such as scouts or general managers, because they choose who gets to be on the teams. The models we have built show statistics that we did not think would directly impact how well their NBA career went, but they did. These statistics include how many three pointers a person attempted, how many free throws they attempted, and how many games they played in. We interpret this as the NBA being successful for people who have tenacity, and are willing to work for their career. The work we did is not only applicable to the NCAA but also to other leagues too. If we could apply this to other people who play in leagues abroad, we could increase opportunities for them to enter the NBA.